基于语义网络的研究兴趣相似性度量方法*

巴志超 [1,2]　李　纲 [1]　朱世伟 [2]
[1](武汉大学信息管理学院　武汉　430072)
[2](山东省科学院情报研究所　济南　250014)

摘要:【目的】为准确识别研究内容相似但使用不同关键词的作者关系，解决传统共现分析方法缺乏语义关联的问题，提出一种基于关键词语义网络构建的作者研究兴趣相似性度量方法。【方法】通过引入 word2vec 模型对作者关键词进行词向量表示，将关键词表示成语义级别的低维实值分布；计算关键词之间的语义相关度并构造关键词语义网络，采用 JS 距离对构建的作者研究兴趣矩阵进行相似性度量。【结果】该方法能计算出共现及非共现词对的相关性，有效地挖掘出作者之间的潜在合作关系。【局限】训练语料的数量和准确性有待进一步提高，提出的度量方法仅考虑两个作者之间的潜在合作关系。【结论】研究结果对改进基于传统的共现分析方法度量作者合作关系具有重要的参考价值。

关键词: 作者关键词网络　神经网络语言模型　语义相似度　研究兴趣矩阵
分类号: G250

1 引　言

有效识别作者研究兴趣的相似度，是挖掘科研人员潜在合作关系以及探测学科知识结构的重要基础工作。针对作者研究兴趣的相似度计算已在学科知识结构探测[1]、科研社区发现[2]、作者合著结构剖析[3]、学科间关系探讨[4]等领域取得广泛的应用。在当前科研工作的大团体中，如何准确地识别作者研究兴趣之间的相似性，有效挖掘潜在的竞争对手与合作伙伴，一直以来也是图书情报领域研究的重要课题。

针对作者研究兴趣相似度计算问题，相关学者已经开展了大量的研究工作。目前，采用的主要方法有以文献为计量单位的作者共被引分析[5]、作者文献耦合分析[6]以及以关键词为计量单位的关键词分析方法[7]等。由于关键词是文献核心内容的浓缩和提炼，高度概括了文献的基本内容，较作者合著和引文分析方法，基于关键词分析更能直观地反映出文献内容和作者的研究兴趣，因而较多的研究利用作者发表的文献资源中的关键词集合来揭示作者的研究兴趣。然而，基于这种词频或共现词频的分析方法，假定作者所使用的关键词之间相互独立，未考虑关键词之间的语义关联信息，因而不能很好地刻画出词之间的相似程度，无法有效地挖掘研究内容相似但使用不同关键词的作者关系[8]。另外，该方法只是直观地假设共现就必然相关，且在共现词数相同的情况下关键词之间的相关强度完全相同。从单篇文献以及领域范围内整个文献集合的研究角度而言，共现的关键词对之间存在直接的共现关系或间接的语义关联，且关联强度不同，而不共现的关键词对之间也存在一定的关联性。

因此，为有效挖掘作者所使用的关键词之间的语义关系，本文在传统作者关键词分析方法的基础上，提出一种基于关键词语义网络构建的作者兴趣相似性度量方法。首先，通过浅层神经网络语言模型 word2vec

通讯作者：巴志超，ORCID：0000-0001-5626-5604，E-mail：bazhichaoty@126.com。

*本文系国家自然科学基金项目“科研团队动态演化规律研究”(项目编号：71273196)、山东省重点研发计划项目“可定制大数据知识服务平台关键技术研究及应用”(项目编号：2015GGX101037)和山东省科学院青年基金项目“基于本体标注的科技文档挖掘方法关键技术研究”(项目编号：2013QN036)的研究成果之一。

对作者文献进行建模学习，将作者的关键词表示成语义级别的单词特征向量，通过 Pearson 相关系数计算关键词之间的相关程度；其次，构造关键词语义矩阵作为作者的研究兴趣矩阵，通过 Jensen-Shannon 距离计算作者之间的相似性；最后，选取国内电子政务研究领域的核心著者作为对象进行实验，验证该方法的有效性。

2 相关工作

基于以文献为计量单位的作者共被引分析、作者文献耦合分析等方法，主要是通过计算作者之间的共被引强度、耦合强度来度量作者研究兴趣的相似程度。如 Jan Van Eck 等[9]分别采用 Pearson 相关系数和 Salton 余弦相似度计算方法从概率分布角度探讨作者共被引相似度的度量。邱均平等[10]从多方面对获取的引文网络进行重构，并引入时间维度来探索引文网络中的知识扩散和演进过程。Zhao 等[11]首次提出作者文献耦合分析方法，并将该方法应用于世界范围内情报学领域的演化研究中。随后，陈远等[12]在国内首次对该方法进行实证应用，用于探索国内情报学领域的前沿性学科结构以及研究热点状况。王知津等[13]将 1990 年–2009 年我国情报学研究进行不同时段的划分，采用该方法识别情报学总体研究领域和各时段的研究领域。上述方法主要通过借助第三方文献而建立一种隐性的、间接的学术关系，在揭示作者研究内容上不如直接以关键词为计量单位的作者关键词共现分析方法[14]。Morris 等[15]也认为通过关键词的共现关联在一起的文献更有可能表达同一个研究主题。另外，基于作者文献耦合分析方法只考虑两个学者之间共同引用参考文献的数量，而未考虑参考文献之间内容上的关联性。

基于文献的关键词共现分析方法，主要通过统计作者所使用的关键词的共现频次来度量作者研究兴趣的相似度。如 Onyancha 等[16]引入社会网络分析中的复杂网络相关理论，利用关键词的共词矩阵分析整体网络特性，并对网络中节点性质和存在的派系进行研究，进一步分析研究领域的知识或主题结构。邱均平等[17]采用作者关键词共现分析方法挖掘我国计量学领域的隐性作者合作关系，并采用 Ucinet 对我国计量学领域的综合性作者合作关系进行可视化分析。丁敬达[18]将作者关键词耦合与作者文献耦合分析方法相结合，用于揭示创新知识社区内部的科学交流特征与规律研究，拓展这两种计量方法的应用视角与应用范围。陈卫静等[14]提出在作者关键词耦合分析方法的基础上，综合考虑关键词的频次、作者发文量及关键词分布等因素对作者相似度的影响，采用 TF-IDF 的关键词加权方式，对关键词耦合强度的计算方法进行改进。然而，上述研究大多是利用作者文献中关键词的共现强度来分析作者之间的关系，采用类似 One-Hot Representation 的建模方式构建共现矩阵，如果共现则该关键词对在共现矩阵中取值为 1，否则为 0，且直观地假设共现就必然存在相关，缺乏对关键词对之间语义关系和关系强度的揭示。为此，本文采用 word2vec 嵌入模型对作者的文献集合进行语义建模学习，从语义和语法的角度计算关键词之间的相关强度，进而计算出作者之间研究兴趣的相似性。由于不共现的关键词对之间也存在一定的关联性，基于该方法也可有效地计算出共现及非共现关键词对之间的相关性。

3 基于语义网络的研究兴趣相似度量方法

3.1 基于 word2vec 模型的语义建模

word2vec 模型是由 Mikolov 等[19]提出用于将单词转化成向量的深度学习工具，与主题模型如 PLSA、LDA 等不同的是，该词嵌入模型主要利用词汇与上下文信息的共现，基于窗口长度考虑语法和语义更底层的信息进行建模，能更有效地刻画出词与词之间的语义关系。word2vec 模型为获取词的向量表示提供两种有效的建模方法：基于连续词袋(Continuous Bag-Of-Words，CBOW)和 Skip-gram 架构。本文主要基于 CBOW 模型，并采用 Hierarchical Softmax 方法进行优化训练。

CBOW 模型主要采用的神经网络框架是在 Hierarchical NNLM 的基础上去掉最耗时的非线性隐藏层，并让输入层的所有单词共享映射层[20]。该模型通过利用单词的上下文信息来生成单词的词向量，并对生成的上下文词向量进行求和得到训练的目标向量，结合词频计算权值构建 Huffman 树，利用异步随机梯度下降的方法对目标函数进行训练，CBOW 模型的框架如图 1 所示。

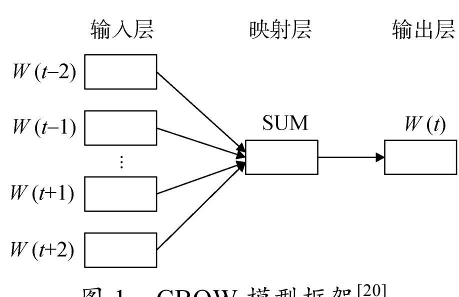


图 1　CBOW 模型框架[20]

基于神经网络的语言模型的目标函数为对数似然函数 $\zeta=\sum_{w\in C}\log p(w|Context(w))$，其中关键是对条件概率函数 p(w|Context(w))的构造。基于 Hierarchical Softmax 优化的 CBOW 模型主要利用词向量 X(w)和 Huffman 树来定义条件概率函数 p(w|Context(w))，定义如下[20]:

$$p(v^w\mid context)=\prod_{j=1}^{L}p(v_j^w\mid v_1^w,v_3^w,\cdots,v_{j-1}^w,context)^{1-v_j}(1-p(v_j^w\mid v_1^w,v_3^w,\cdots,v_{j-1}^w,context))^{v_j} \quad (1)$$

其中，$v^w=(v_1^w,v_2^w,\cdots,v_L^w)\in(0,1)$表示当前词 w 的哈夫曼编码，$(w_{t-n+1},\cdots,w_{t-1},w_{t+1},\cdots,w_{t+n-1})$表示词 w 的上下文，简记为 context。通过 CBOW 模型得到关于词指定长度的向量，使用这组向量采用余弦值或欧式距离来计算词语之间的语义相似度。

3.2 作者研究兴趣的表示

通过建模得到的关键词集构建作者研究兴趣表示模型时，通常会将每个作者的研究内容或兴趣表示为 $\chi_i=\{(k_1, W_{i1}), (k_2, W_{i2}), \cdots, (k_m, W_{im})\}$的形式，其中$\{k_1, k_2, \cdots, k_m\}$表示为作者 χ_i 发表的文献中所使用的关键词集合，$\{W_{i1}, W_{i2}, \cdots, W_{im}\}$表示作者所使用的关键词出现的词频或频率值，未出现则计为 0 值。由于基于这种共现方法假定关键词之间相互独立，无法有效地获取关键词之间的相关程度，而作者所使用的关键词之间也是存在关联关系的，对于采用相似关键词的作者而言，其研究内容或兴趣也存在一定的相似性。

为此，本文通过 word2vec 嵌入模型对所有作者发表的文献中的题名及摘要信息进行建模学习，通过将每个作者所使用的关键词转化为语义级别的单词特征向量形式，然后再计算关键词之间的语义相似度。对于作者 χ_i 的关键词 k_{ij}，可将关键词 k_{ij} 表示为：$k_{ij}=\{(k_1, S_1), (k_2, S_2), \cdots, (k_n, S_n)\}$的特征向量形式，其中$\{k_1, k_2, \cdots, k_n\}$表示与词 k_{ij} 最相关的 n 个词语，$\{S_1, S_2, \cdots, S_n\}$表示各词语与关键词 k_{ij} 之间的余弦距离值。如将作者关键词"电子政务"表示为向量：{(电子政府: 0.848), (公共服务: 0.825), (政府网站: 0.751), (服务型政府: 0.731), (信息服务: 0.712)}。

在获得每个关键词的向量表示后，采用 Pearson 相关系数(Pearson Correlation Coefficient, PCC)[21]计算关键词之间的语义相关性。Pearson 相关系数常用于度量两个随机变量 X 与 Y 之间的线性相关性，通过利用 Pearson 相关系数，可得到关键词 k_i 的词向量 $S_{i,k}$ 与关键词 k_j 的词向量 $S_{j,k}$ 之间的相关性，计算公式[21]如下:

$$\rho(k_i,k_j)=\frac{\sum_k(S_{i,k}-\bar{S}_i)(S_{j,k}-\bar{S}_j)}{\sqrt{\sum_k(S_{i,k}-\bar{S}_i)^2}\sqrt{\sum_k(S_{j,k}-\bar{S}_j)^2}} \quad (2)$$

其中，$\bar{S}_i$、$\bar{S}_j$ 表示关键词 k_i、k_j 与其所有相关词之间余弦的平均值。计算得到的 $\rho(k_i, k_j)$值越大，说明关键词之间越相关。

3.3 关键词的语义网络构建

获得作者表示模型后，需要构建作者–关键词网络。基于传统的关键词共现分析方法通常构建词共现矩阵对作者关系进行量化计算，然而，构建的二值/多值矩阵中统计的原始词对频次是绝对值，难以反映词与词之间真正的相互依赖程度。同时，多值矩阵中存在的频次悬殊数据以及较多的零值会对最终的统计结果造成影响。为此，相关学者提出采用关键词共现指数表达的方法，通过引入关键词共现相对强度指标对词对频次进行包容化处理，生成相似矩阵和相异矩阵。如采用 E 指数[22]、Ochiia 系数[23]等。这几种方法只是为减少低频词对共词分析过程的干扰，以区分对待低频词以及高频词之间的共现强度，但仍无法挖掘出关键词之间的语义关联信息。

本文提出通过公式(2)计算得到的关键词之间的语义相关度作为矩阵的元素值，将传统的词共现矩阵转化为元素值在[0, 1]区间取值的相关矩阵形式，最终可构建关键词的语义网络 G_s 如下:

$$G_S=\begin{bmatrix}\rho(k_1,k_1) & \rho(k_1,k_2) & \cdots & \rho(k_1,k_n)\\ \rho(k_2,k_1) & \rho(k_2,k_2) & \cdots & \rho(k_2,k_n)\\ \vdots & \vdots & & \vdots\\ \rho(k_n,k_1) & \rho(k_n,k_2) & \cdots & \rho(k_n,k_n)\end{bmatrix} \quad (3)$$

其中，矩阵元素 $\rho(k_i, k_j)$表示为关键词 k_i, k_j 之间的

相关度，该数值越大，表明关键词 k_i，k_j 之间的关联程度就越强。若一个关键词与较多的关键词之间计算得到的相关度都较高，则说明该关键词对表示作者研究兴趣的重要程度较高，越能代表该作者的研究兴趣。基于这种通过作者文献集学习得到的关键词向量形式，能够较好地表达作者的研究兴趣。

3.4 作者研究兴趣相似性计算

通过以上步骤可将作者的研究兴趣表示成兴趣矩阵 $G_s^i(m\times n)$形式，矩阵中行表示关键词 k_i 的词向量，列表示作者所使用的 m 个关键词。在计算作者研究兴趣的相似性时，只需要计算两个研究兴趣矩阵之间的语义关系。本文采用兴趣矩阵之间的 Jensen-Shannon 距离(Jensen-Shannon Divergence，JSD)[24]作为研究兴趣的相似性度量。JSD 广泛用于计算两个概率分布之间的相似度。具体地，对于两个离散概率分布 P 和 Q，它们之间的 Jensen-Shannon[24]可被定义为:

$$JSD(P\|Q)=\frac{1}{2}(KL(P\|M)+KL(Q\|M)) \tag{4}$$

其中，M=(P+Q)/2，KL(· ‖ ·)表示两个分布之间的 Kullback-Leibler 距离。通过采用 Jensen-Shannon 距离作为关键词之间的相似度度量，可将作者 χ_i 所使用的关键词 k_{ij} 与作者 χ_j 所使用的关键词 k_{jt} 之间的相似度 $S_{k_{ij},k_{jt}}$ 定义为:

$$S_{k_{ij},k_{jt}}=\frac{1}{JSD(p(D|k=k_{ij})\|p(D|k=k_{jt}))} \tag{5}$$

基于公式(5)获得关键词向量两两之间的相似度 $S_{k_{ij},k_{jt}}$ 后，取所有关键词向量之间相似度的平均值作为作者研究兴趣矩阵之间的相似度。最终得到作者 χ_i 的兴趣矩阵 G_s^i 与作者 χ_j 的兴趣矩阵 G_s^j 之间的相似度 $S_{G_s^i,G_s^j}$ 为:

$$S_{G_s^i,G_s^j}=\frac{1}{m^2}\sum_{j=1}^{m}\sum_{t=1}^{m}S_{k_{ij},k_{jt}} \tag{6}$$

可知，$S_{G_s^i,G_s^j}$ 越大，说明两个作者的关键词之间的 JSD 越小，作者更倾向关联。

4 实证分析

4.1 实验设置

选取《中文社会科学引文索引》(CSSCI)作为数据源，获取国内电子政务研究领域的期刊文献。以(主题="电子政务"or 主题="移动政务"or 主题="电子政府"or 主题="政府网站"or 主题="政务微博")为检索式，发表在 2003 年–2014 年时间段内的文献共 2 956 篇，去除综述、评论、报告及其他类型文献，并根据文献的标题及发表年份进行去重处理，最终获得期刊论文共 2 791 篇，涉及作者 2 104 位，关键词 4 725 个。

根据普赖斯理论，发表论文数为 N 篇及以上的作者为该研究领域的核心作者，N=0.749 $(\eta_{max})^{1/2}$，其中 η_{max} 表示该研究领域发文数量最多的作者的论文数。选择发文量为 6 篇及其以上的作者(共 51 位)作为本研究的分析对象，挖掘这些核心作者的研究兴趣的相似性。通过对 51 位核心作者所使用的关键词进行抽取与词频统计，发文量最高作者所使用的关键词数量为 86 个，而发文量最低作者所使用的关键词数量为 21 个。为便于作者兴趣矩阵相似性的计算，本文针对各核心作者选取相同数量的关键词进行 word2vec 建模学习。另外，在选取关键词表示作者研究兴趣时，删除对分析作者研究兴趣相似性以及分析领域热点较低贡献的概括性关键词，如电子政务、电子政府等。

4.2 实验结果与分析

(1) 作者关键词语义建模

对作者的关键词向量进行建模，根据下载的题录信息，对所有作者文献中的题名和摘要进行分词和停用词过滤，而对作者自标引的关键词不进行分词处理。获取分词的结果后，采用 word2vec 嵌入模型进行训练学习，生成词向量库文件，获取每个作者关键词的词向量形式。对于模型的训练参数设定为：最相似词维度 topNSize=40，上下文窗口大小参数 window=5，设定高频词亚采样阈值参数 sample=1e-3，并采用层次 Softmax 和 CBOW 算法 hs=1、cbow=1。表 1 显示了部分作者关键词的词向量。可以看出，基于 word2vec 模型对作者题名和摘要进行建模，将作者关键词表示成词向量形式，能够有效地挖掘出与作者关键词比较相关的词。关于"公共服务"主题，相关作者主要从政府职能、社会服务及信息化等方面进行研究，关于"政府微博"主题，主要是从政府管理、网络监管、公共事务及信息传播等方面进行研究，这与获取的作者文献的研究内容基本一致。

表 1　作者关键词的 Top-6 相关词向量(部分)

关键词	Top-6 相关词向量		
公共服务	政府: 0.732	职能: 0.711	服务: 0.650
	社会管理: 0.636	信息化: 0.582	公共: 0.535
政府微博	政府管理: 0.712	网络治理: 0.701	监理: 0.632
	公共事务: 0.621	隐私: 0.579	政务: 0.523
信息资源	信息共享: 0.683	整合: 0.663	信息化: 0.641
	信息服务: 0.622	分散: 0.579	资源: 0.420
绩效评估	定量: 0.724	评价: 0.702	立法: 0.683
	价值取向: 0.647	服务政府: 0.604	改进: 0.601
云计算	个性化: 0.762	信息服务: 0.738	异构: 0.694
	统筹规划: 0.632	效用: 0.614	技术: 0.602

(2) 关键词语义网络分析

在基于词频或共现词频的共现分析方法中，只是根据作者自标引的关键词对是否在同一篇文献中共现来确定关键词对在共现矩阵中的取值，如图 2 所示。关键词“信息资源”与“政府微博”之间不存在相关性，与“绩效评估”只共现 2 次；而关键词“公共服务”与“政府微博”、“公共管理”具有较高的相关性。在这种多值矩阵中存在的频次悬殊数据以及较多的零值也严重影响最终的共词分析结果。通过提出的基于 word2vec 模型的语义相关度量方法，可有效计算出共现及非共现关键词对之间的关系，计算得到的关键词之间具有不同程度的相关性，一定程度上提高了共词分析方法的科学性和有效性，结果如图 3 所示。

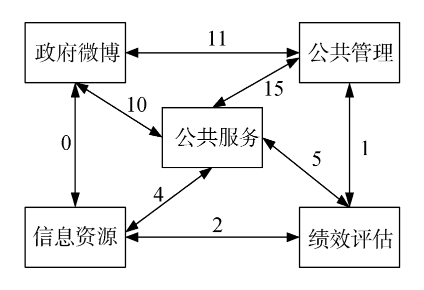


图 2　关键词频次共现度量关系

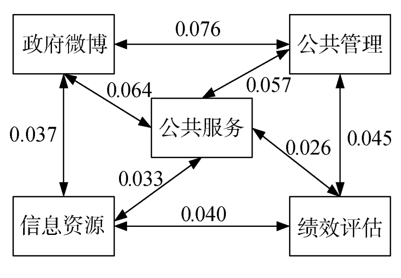


图 3　关键词语义关系

为进一步研究作者所使用关键词之间的语义关系以及对作者研究兴趣的贡献程度，对构建的关键词语义网络进行分析。为突出作者所使用关键词之间的语义相关性，设定阈值 λ=0.05，小于该值计为 0 值，如图 4 所示：

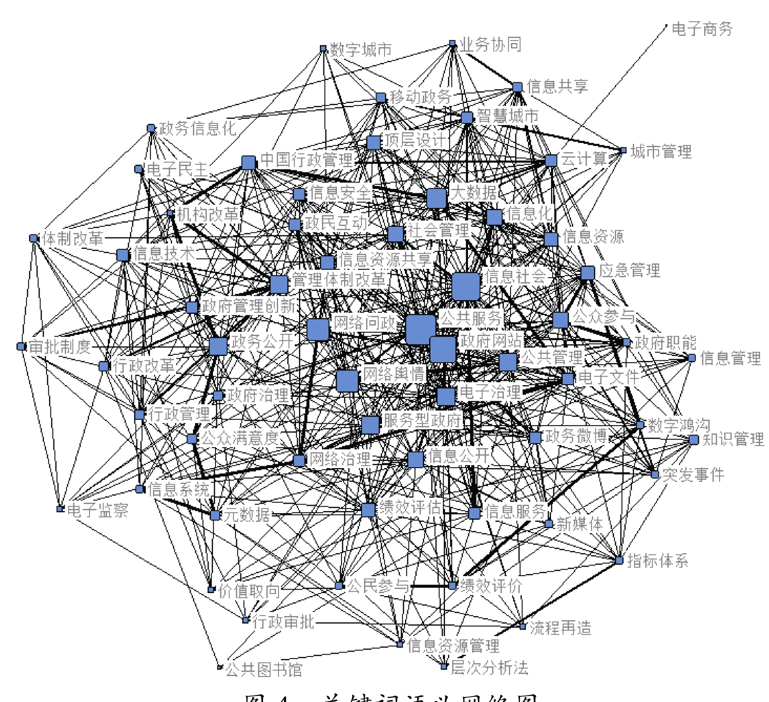


图 4　关键词语义网络图

其中，节点的位置越居中并且面积越大，说明该作者关键词越核心，对作者研究兴趣贡献度越大。可以看出，“公共服务”、“信息社会”、“政府网站”和“网络舆情”等关键词所对应的节点较大且处于中间位置，说明这些关键词为该领域研究的主要内容，受到相关研究者的广泛关注。而“信息管理”、“政府职能”、“指标体系”、“层次分析法”等关键词所对应的节点较小且处于相对边缘位置，说明这些方面的研究已趋近于饱和。另外，关键词之间连线的粗细程度也可以看出关键词之间的语义相关强度，“公共服务”、“政府网站”、“信息服务”等关键词与其他关键词相关性都比较大，说明这些关键词对作者研究兴趣也具有较大的贡献程度。

(3) 作者研究兴趣相似性分析

在构建关键词语义网络后，将每个作者的研究兴趣表示成研究兴趣矩阵形式，采用公式(6)计算作者研究兴趣的相似度，并进行归一化处理。通过归一化处理将矩阵对角线元素设置为1，以突出作者与自身的相似性，表2显示作者研究兴趣之间的相似程度。本文提出的作者兴趣矩阵相似度计算方法仅研究两个作者之间潜在的合作关系，通过该方法可以计算出每对作者之间研究的相似程度，能够有效挖掘出相似度较大但尚未产生合作关系的作者对。如对于作者罗贤春，其与张锐昕、何振、孟庆国等作者研究兴趣比较相似，而作者王芳与郑磊、高洁等作者具有较高的相似性。基于该方法计算同单位且经常产生合作关系的作者之间也具有较高的相似度，如计算得到同属吉林大学的作者张锐昕与杨国栋之间的相似度为0.372，同属湘潭大学的作者何振与周伟之间的相似度为0.410，说明该方法在计算作者研究兴趣相似性时具有一定的有效性。

表 2 作者研究兴趣相似矩阵(部分)

作者＼作者	罗贤春	张锐昕	何振	王芳
罗贤春	1			
张锐昕	0.315	1		
何　振	0.472	0.336	1	
王　芳	0.190	0.216	0.185	1
刘焕成	0.157	0.193	0.215	0.206
胡广伟	0.092	0.107	0.143	0.253
郑　磊	0.231	0.214	0.195	0.421
徐晓林	0.074	0.095	0.130	0.206
孟庆国	0.321	0.284	0.264	0.175
高　洁	0.145	0.172	0.193	0.341

为进一步挖掘作者研究兴趣之间的语义关系，将作者作为节点进行社会网络分析，选择相似度大于0.20 的作者关系进行可视化呈现，如图 5 所示：

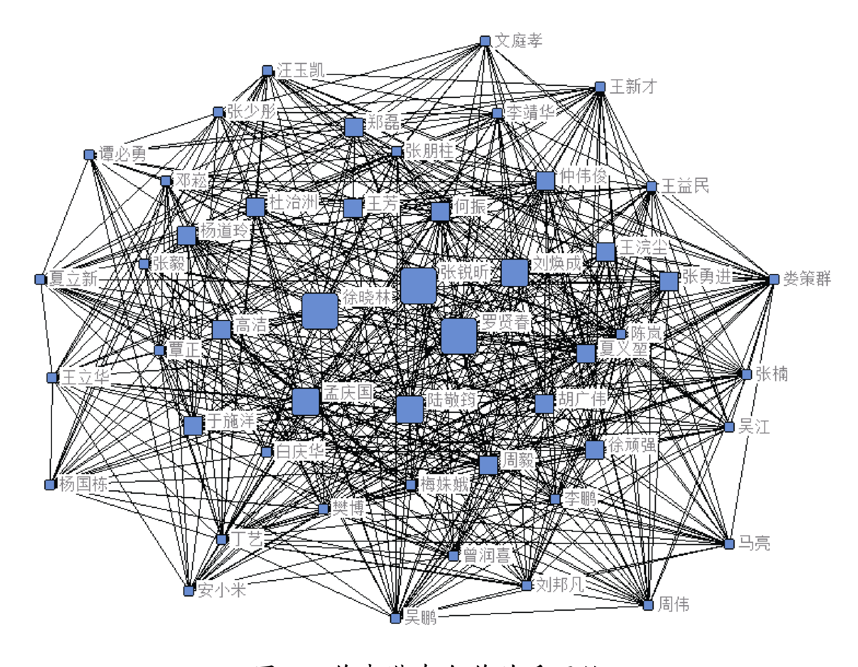


图 5 作者潜在合作关系网络

可以看出，罗贤春、张锐昕、刘焕成等作者处于网络的核心位置，说明这些作者在该领域具有较高的影响，而马亮、安小米、王新才等处于网络的边缘位置，说明这些作者的影响力相对较小。对该网络进行中心性分析，计算得到该网络的平均点度中心度为 46.088，其中有 17 位作者的点度中心度大于平均节点中心度值，且 17 位作者中有 8 位位于作者发文量统计前 10 名，说明该方法计算得到的点度中心度和作者发文量存在相关关系且之间的拟合程度较高。其中，夏义堃、胡广伟、张锐昕等点度中心度比较高，说明这些作者与较多作者的研究兴趣具有较高相似度，在该领域研究比较广泛且具有较高影响力。罗贤春、张锐昕、何振三位作者的中间中心度最高，且与其他作者存在一定的差距，说明三位作者处于重要的地位，且较大影响其他作者之间研究兴趣的相似性，更多的作者之间可通过这三位作者取得较高的研究兴趣相似度，从而建立潜在的合作关系。通过该方法将研究兴趣相似但未产生合作关系的作者进行关联，从而为更多具有相同研究兴趣的作者之间进行知识的交流提供借鉴，以促进电子政务领域中热门研究主题的发展。

(4) 作者研究兴趣类团主题分析

通过聚类算法将研究兴趣相似的作者进行聚集形成类团，以揭示该领域的研究主题结构。对作者研究兴趣相似矩阵进行层次聚类分析，得到国内电子政务研究领域的作者聚类树状图，如图 6 所示。可以看出，对作者研究主题进行类团分析，可将该网络大致分为三个类团。类团 1 中主要包含罗贤春、何振、张锐昕等作者，该类团研究方向比较多元化，主要是关于电子政务建设、体系构建、信息服务等比较宏观理论的研究，以探索电子政务相关的理论框架，同时还涉及信息共享、科学决策、绩效评估等研究主题。类团 2 中主要包含刘焕成、胡广伟、高洁等作者，研究内容主要从信息管理、体制改革、信息技术等角度探索电子政务领域的信息整合、重组与管理等微观策略方面的研究，还涉及到政府职能、公共管理、行政管理等应用技术的研究。类团 3 主要包括王芳、郑磊、徐晓林，三位作者共有关键词为政务微博、网络伦理、移动政务等，即围绕以政府微博为主题的网络舆情、电子服务开展研究。从聚类结果可看出，基于构建的作者研究兴趣矩阵聚集形成的类团语义都比较明确。

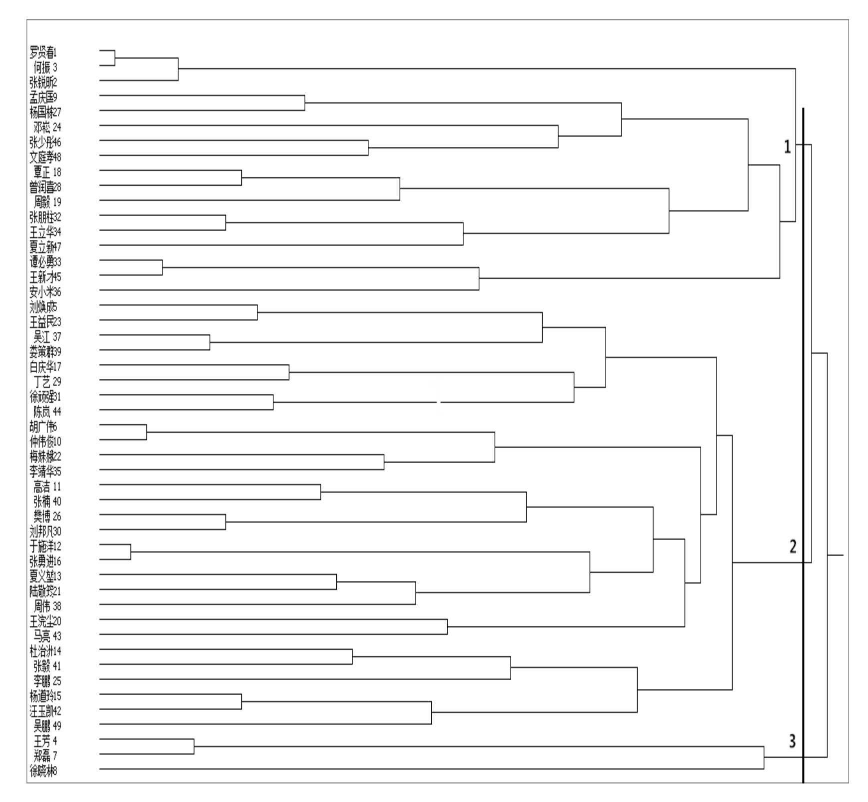


图 6　电子政务领域作者聚类树状图

5 结 语

本文提出一种基于关键词语义网络构建的作者研究兴趣相似性度量方法。通过将深度学习思想引入到文献计量中，基于word2vec模型对作者文献题名及摘要进行建模，将作者所使用的关键词表示成语义级别的单词特征向量，从而将作者研究兴趣表示成矩阵形式进行相似性度量。通过对国内电子政务研究领域的核心作者进行分析，验证了该方法能够有效地挖掘作者研究兴趣之间的相似性。对作者研究兴趣相似性的有效度量，能够帮助学者选择与自己研究兴趣相似但还未产生合作关系的学者进行知识的交流提供借鉴。

本文的不足之处在于：只采用作者文献的题名和摘要进行建模学习，训练语料的准确性有待进一步提高。训练时所使用的语料越准确、全面，建模得到的词向量越相关，下一步的研究工作将获取文献的摘要和正文对关键词进行训练学习。本文提出的度量方法仅计算两个作者之间潜在的合作关系，将作者研究兴趣之间的相似性转化成兴趣矩阵之间的相似性度量。需要提出新的有效度量方法计算多个作者之间的相似程度，以挖掘多个作者之间共同合作的可能性。只是借助词模型进行关键词对间的语义度量，可借助外部知识库 Wikipedia、HowNet 词典等，构建更加丰富的科研合作网络、引用关系网络并选择合适的评价指标进行验证和分析，以期进一步提高共词分析方法的有效性。

参考文献：

[1] 邱均平，刘国徽，董克. 基于合作分析的知识聚合与学科知识结构研究——以国内知识管理领域为例[J]. 情报理论与实践，2014, 37(8): 6-11. (Qiu Junping, Liu Guohui, Dong Ke. Research on Knowledge Aggregation and Discipline Structure Based on Collaboration Analysis—Taking the Field of Knowledge Management in Domestic as an Example [J]. Information Studies: Theory&Application, 2014, 37(8): 6-11.)

[2] 李纲，李岚凤，毛进，等. 作者合著网络中研究兴趣相似性实证研究[J]. 图书情报工作，2015, 59(2): 75-81. (Li Gang, Li Lanfeng, Mao Jin, et al. Empirical Research on Similarity of Research Interests in Co-authorship Network [J]. Library and Information Service, 2015, 59(2): 75-81.)

[3] 王福生，石秀春，杨洪勇. 基于作者簇的科研合作网络模型[J]. 情报理论与实践, 2009, 32(1): 35-37. (Wang Fusheng, Shi Xiuchun, Yang Hongyong. Research on Scientific Collaboration Network Based on Author Cliques [J]. Information Studies: Theory & Application, 2009, 32(1): 35-37.)

[4] Abramo G, D'Angelo C A, Costa F. Identifying Interdisciplinary Through the Disciplinary Classification of Coauthors of Scientific Publications [J]. Journal of the American Society for Information Science and Technology, 2012, 63(11): 2206-2222.

[5] 邱均平，张晓培. 基于 CSSCI 的国内知识管理领域作者共被引分析[J]. 情报科学，2011, 29(10): 1141-1145. (Qiu Junping, Zhang Xiaopei. Author Co-citation Analysis of Knowledge Management in China Based on the CSSCI [J]. Information Science, 2011, 29(10): 1141-1145.)

[6] 宋艳辉，武夷山. 基于作者文献耦合分析的情报学知识结构研究[J]. 图书情报工作，2014, 58(1): 117-123. (Song Yanhui, Wu Yishan. Resarch on Knowledge Structure of Information Science Based on Author Bibliographic-coupling Analysis [J]. Library and Information Service, 2014, 58(1): 117-123.)

[7] 孙海生. 作者关键词共现网络及实证研究[J]. 情报杂志，2012, 31(9): 63-67. (Sun Haisheng. Author Keyword Co-Occurrence Network Analysis: An Empirical Research [J]. Journal of Intelligence, 2012, 31(9): 63-67.)

[8] 刘萍，郭月培，郭怡婷. 利用作者关键词网络探测作者相似性[J]. 现代图书情报技术，2013(12): 62-69. (Liu Ping, Guo Yuepei, Guo Yiting. Use of Author-Keyword Network for Detecting Author Similarity [J]. New Technology of Library and Information, 2013(12): 62-69.)

[9] Jan Van Eck N, Waltman L. Appropritate Similarity Measure for Author Co-citation Analysis [J]. Journal of the American Society for Information Science and Technology, 2008, 59(10): 1653-1661.

[10] 邱均平，李小涛. 基于引文网络挖掘和时序分析的知识扩散研究[J]. 情报理论与实践，2014, 37(7): 5-10. (Qiu Junping, Li Xiaotao. Research on Knowledge Diffusion Based on Citation Network Mining and Timing Analysis [J]. Information Studies: Theory & Application, 2014, 37(7): 5-10.)

[11] Zhao D, Strotman A. Evolution of Research Activities and Intellectual Influences in Information Science 1996-2005: Introducing Author Bibliographic-coupling Analysis [J]. Journal of the American Society for Information Science and Technology, 2008, 59(13): 2070-2086.

[12] 陈远，王菲菲. 基于 CSSCI 的国内情报学领域作者文献耦

合分析[J]. 情报资料工作, 2011, 32(5): 6-12. (Chen Yuan, Wang Feifei. An Analysis on the Bibliographic Coupling in the Field of Information Studies in China: Based on CSSCI [J]. Information and Documentation Services, 2011, 32(5): 6-12.)

[13] 王知津, 周鹏, 谢丽娜. 用 ABCA 方法识别和阐释我国当代情报学研究领域[J]. 情报学报, 2013, 32(1): 4-12. (Wang Zhijin, Zhou Peng, Xie Lina. The Identification and Explanation of Research Fields of Contemporary Information Science in China Using ABCA Method [J]. Journal of the China Society for Scientific and Technical Information, 2013, 32(1): 4-12.)

[14] 陈卫静, 郑颖. 基于作者关键词耦合的潜在合作关系挖掘[J]. 情报杂志, 2013, 32(5): 127-131. (Chen Weijing, Zheng Ying. Mining Potential Cooperative Relationships Based on the Author Keyword Coupling Analysis [J]. Journal of Intelligence, 2013, 32(5): 127-131.)

[15] Morris S A, Yen G G. Crossmaps: Visualization of Overlapping Relationships in Collections of Journal Papers [J]. Proceedings of the National Academy of Sciences, 2004, 101(S1): 5291-5296.)

[16] Onyancha O B, Ocholla D N. Is HTV/AIDS in Africa Distinct? What Can We Learn from an Analysis of the Literature [J]. Scientometrics, 2009, 79(1): 277-296.

[17] 邱均平, 陈木佩. 我国计量学领域作者合作关系研究[J]. 情报理论与实践, 2012, 35(11): 56-60. (Qiu Junping, Chen Mupei. Research on Author Collaboration in the Metrology Field in China [J]. Information Studies: Theory&Application, 2012, 35(11): 56-60.)

[18] 丁敬达. 创新知识社区内部科学交流的特征和规律——基于某国家重点实验室的实证分析[J]. 情报学报, 2011, 30(10): 1086-1094. (Ding Jingda. Characteristics and Regularity in Scientific Communication Within Innovative Knowledge Community: An Empirical Study of a State Key Laboratory [J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(10): 1086-1094.)

[19] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality [C]. In: Proceedings of the Neural Infornational Processing Systems Conference. Nevada, United States: Neural Information Processing Systems Foundation, 2013: 3111-3119.)

[20] Morin F, Bengio Y. Hierarchical Probabilistic Neural Network Language Model [C]. In: Proceedings of the International Workshop on Artificial Intelligence and Statistics. Cambridge: Cambridge University Press, 2005: 246-252.

[21] Polzehl J, Spokoiny V. Propagation-Separation Approach for Local Likelihood Estimation [J]. Probability Theory and Related Fields, 2006, 135(3): 335-362.

[22] Callon M, Courtial J P, Laville F. Co-word Analysis as a Tool for Describing the Network of Interactions Between Basic and Technological Research: The Case of Polymer Chemsitry [J]. Scientmetrics, 1991, 22(1): 155-205.

[23] 郑华川, 于晓欧, 辛彦. 利用共词聚类分析探讨抗原 CD44 研究现状[J]. 中华医学图书情报杂志, 2002, 11(2): 1-3. (Zheng Huachuan, Yu Xiaoou, Xin Yan. Antigen CD44 with Clustered Analysis of Co-words: A Status Quo Investigation [J]. Chinese Journal of Medical Library and Information Science, 2002, 11(2): 1-3.)

[24] Endres D M, Schindelin J E. A New Metric for Probability Distributions [J]. IEEE Transactions on Information Theory, 2003, 49(7): 1858-1860.

作者贡献声明:

巴志超, 李纲, 朱世伟: 提出研究思路, 设计研究方案, 论文最终版本修订;
巴志超: 进行实验, 采集、清洗和分析数据, 起草论文。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 巴志超, 李纲, 朱世伟. ele_govdata.txt. 电子政务研究领域的期刊文献实验数据.

[2] 巴志超, 李纲, 朱世伟. word2vec.rar. word2vec 建模程序包.

[3] 巴志超, 李纲, 朱世伟. CorMatrix.xls. Top-100 关键词共现词频矩阵.

[4] 巴志超, 李纲, 朱世伟. EquCorMatrix.xls. 基于 Equivalence 系数的相似矩阵.

[5] 巴志超, 李纲, 朱世伟. AuthorSim.xls. 作者相关矩阵和作者–关键词矩阵.

[6] 巴志超, 李纲, 朱世伟. Data_XML.html. 文献题录转化格式文档.

[7] 巴志超, 李纲, 朱世伟. AuthorSimi.##h. 用于类团分析的作者相关矩阵格式.

收稿日期: 2015-12-02
收修改稿日期: 2015-12-28

Similarity Measurement of Research Interests in Semantic Network

Ba Zhichao[1, 2] Li Gang[1] Zhu Shiwei[2]
[1](School of Information Management, Wuhan University, Wuhan 430072, China)
[2](Information Research Institute of Shandong Academy of Sciences, Ji'nan 250014, China)

Abstract: [**Objective**] This study aims to identify relationship among authors of papers with similar contents but different keywords, and then tries to add more sematic factors to the co-occurrence analysis. [**Methods**] We proposed a method to gauge the similarity of research interests based on the keywords semantic network system. First, all keywords were represented as word vectors and translated into low dismension distribution with the help of neural network language—word2vec model. Second, we calculated the semantic association of keywords to build up a semantic network. Finally, we adopted the Jensen-Shannon distance method to measure the similarity of research interests. [**Results**] The proposed approach can accurately identify the similarities of co-occurrence and non co-occurrence terms and then effectively predict potential cooperation among authors. [**Limitations**] The amount and accuracy of training materials need to be increased. At present, we could only find potential cooperation between two authors. More research is needed to explore the possibilities of cooperation among multi-authors. [**Conclusions**] The proposed method could help to improve the performance of traditional co-occurrence analysis.

Keywords: Author-keyword network Neural network language model Semantic similarity Matrix of research interests

---

ProQuest 学位论文数据库为美国研究项目提供关键信息

ProQuest 正通过其领先的全球博硕士论文数据库(PQDT)为美国的研究项目提供关键支持。在一个新项目下, ProQuest 通过从数据库中提取重要的提示信息以帮助研究人员更好完成他们的项目。

ProQuest 为 UMETRICS 项目提供数据, 该项目是机构合作委员会的试点项目。这一研究由来自于密歇根大学、俄亥俄州立大学、乔治亚州立大学, 纽约大学和美国人口普查局的研究人员共同完成, 研究成果发表在《科学》杂志上。该研究考察了具有博士学位的人的就业和收入情况。基于一些受资助的研究, 并结合来源于大学、ProQuest 和美国人口普查局的数据, 研究人员随后跟踪并且跨领域的研究博士们的所处位置和收入情况。

ProQuest 的学术交流和论文出版部主任 Austin McLean 表示, " PQDT 是唯一一个综合性的、经过精心维护的, 能提供关键研究数据一站式服务的博硕士论文全文数据库。PQDT 能够为用户提供经过裁剪的数据, 这些定制化的数据为跨学科的研究项目提供支持, 确保研究人员最大化地发掘所拥有信息资源。"

(编译自: http://www.proquest.com/about/news/2016/ProQuest-Dissertation-Database-Provides-Critical-Information.html)

(本刊讯)